DOCUMENT RESUME

ED 184 769 RC 011 948

AUTHOR Chabran, Richard
TITLE Thesaurus Considerations for a National Hispanic Data Base. A Report.
PUB DATE 29 Mar 79
NOTE 33p.; Best copy available

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Bases; *Guidelines; *Hispanic Americans; Indexing; Information Retrieval; Information Systems; *Program Design; Program Development; Reference Materials; *Subject Index Terms; *Thesauri
IDENTIFIERS ERIC; Hispanic Bibliography Info Storage Retrieval Proj

ABSTRACT

Monographs, periodical literature, documents, unpublished reports, and non-print materials in both English and Spanish would be included in a proposed Hispanic data base built around an initial 15 contributing collections totalling an estimated 100,000 items. Because users would employ indexing terms to access the data base, a thesaurus is essential for vocabulary control. Thesaurus guidelines are needed to control syntactical, semantic, and generic problems. The thesaurus, produced and updated by computer, would be organized in four sections: a principal alphabetical display of all terms supported by a rotated display, a hierarchical display, and a descriptor group display. Thesaurus updating, use, and personnel training should be managed in a manner similar to that of ERIC. Instead of creating an entirely new thesaurus or accepting unchanged one of six relevant existing thesauri, the Chicano Thesaurus should serve as a baseline document and certain areas should be developed in greater detail. A list of tasks leading to revision of the thesaurus is included, along with a bibliography. (SB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Draft

ED184769

THESAURUS CONSIDERATIONS

FOR A

NATIONAL HISPANIC DATA BASE

BEST COPY AVAILABLE

A REPORT

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Richard Chabran

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

MARCH 29, 1979

RICHARD CHABRAN
HISPANIC BIBLIOGRAPHY AND INFORMATION
STORAGE AND RETRIEVAL PROJECT

RC 011948

JUL 1979
ERIC/CRESS
RECEIVED

TABLE OF CONTENTS

		Page
A.	System Design	1
B.	Information System Considerations	1
C.	Vocabulary Management	2
D.	Presentation of the Thesaurus	10
E.	Production of the Thesaurus	12
F.	Evaluation	13
G.	Updating	13
H.	Guidelines for use of the thesauri	13
I.	Training	14
J.	Thesauri and Hispanic Studies	14
K.	New Thesaurus versus Adaption	15
L.	Related Thesauri	19
M.	Selected Bibliography	20

Appendices

Thesaurus Considerations for a
National Hispanic Data Base

A. System Design

The actual system design for a Hispanic Data Base will be completed after analysis of returned questionnaires from Hispanic collections that are currently being surveyed. However many aspects of the system can be inferred at this juncture.

B. Information System Considerations

B1 The scope of the information system will be influenced by two major factors, they are: 1) number of contributing collections; 2) the subject areas covered by these collections.

B1.1 Collections. At this point, discussions have centered around having approximately fifteen participating collections/projects from which to build a Hispanic data base. The reasoning behind this is to allow a small number of individuals to work through the problems of "getting started". It can also safely be assumed that the majority of information can be covered if the fifteen collections/projects are well chosen.

B1.2 Scope. The area of Hispanic studies is an interdisciplinary area with the bulk of the information centered in the social sciences, humanities, history and policy studies. Although there are many ways that a data base can be built, discussions within the Hispanic Bibliography Project have centered around the development of clearinghouses which would be responsible for specific subject areas, i.e. mental health, education etc. (Similar to the way that the ERIC system is structured). This should keep duplication to a minimum.

old plantation records

B1.3 The types of literature which are candidates at this point are monographs, periodical literature, documents, unpublished reports, and non print materials.

B1.4 The format for this material is bound volumes, films, tape (both audio, video and magnetic), manuscripts, transparencies and articles/pamphlets.

B1.5 Contents of periodical literature include: articles, book reviews, state of the art reviews, conference reports, creative works, interviews, graphics, editorials, book lists, current events, reports.

B1.6 The language of materials to be incorporated is approximately 80% English, 20% Spanish.

B1.7 The quantity of literature is estimated at 100,000 items. Since 1960 the rate of growth is geographically ~~geographically~~ metrically expanding (Padilla et al).

B1.8 The information system will be used by both experienced users and intermediaries (information specialists, librarians).

B1.9 The number and frequency of users are difficult to estimate, however, the directories listed below point to a set of possible users:

Chicano Affairs Program

Guide to Chicano Studies Departments, Programs, and Centers

National Directory of Chicano Faculty and Research

Additionally, policy planners, students and government officials and general public would all be able to utilize information in the data base.

B1.10 The information system should be able to accommodate both general and specific searches.

B1.11 In most library collections the index terms assigned to a specific item are brought together at the time of indexing i.e. Mexico-History-1848. The user does no coordinating of terms which describe the request, but searches for the most appropriate pre-coorinated terms. In most computer based systems such as ERIC, the indexer assigns many terms (single or multiword) but does not coordinate the terms which best describe the request (thus the system is called post-coordinate indexing) (Kocalis). This system of indexing is more flexable as well as more indepth. This system frequently uses thesauri to index the material.

B1.12 The resources in terms of personnel, staff, equipment and supplies for the development of a Hispanic Data Base will be substantial. It is assumed that participating collections/ projects will have to assume some of this expense but that the initial investment must be funded by an outside funding source.

C. Vocabulary Management

C1 The purpose of vocabulary management is stated concisely by Wall: "Any information storage and retrieval system is, in effect, a communication channel, used to transmit information from its originators to its users. The 'signals' that pass through this communication channel consist only of words; ...Accordingly, in the design and operation of an information system, careful consideration must be given to the vocabulary employed and to the management thereof." Lancaster also points out "...The index language affects performance (of the retrieval operation) at two major points. It affects the search strategy by establishing how precisely the searcher can describe the subject matter of documents...The index language exists primarily to bring the vocabulary of the indexer and the vocabulary of the searcher into coincidence." (See attachment A)

C1.1 Meincke & Atherton point out the state of the theoretical development of vocabulary control "...the lack of conceptual framework which could provide a suitable ordered structure and assist in the information exchange between the structure and the user..."

C1.2 Despite the lack of a well developed theory in the area of vocabulary management, several issues have been identified by Wall:

- .syntactical problems
- .sematic problems
- .generic problems
- .viewpoint problems

C2 Review of Techniques

In a review of techniques of vocabulary development and control (see attachment B) Wall states: "the most sophisticated and the most capable--and also the most expensive--vocabulary control device is the information retrieval thesaurus... it is capable of providing excellent solutions to all the basic problems of verbal communication."

C3 Construction of Thesauri

The basic issues of controlled vocabulary will be discussed in relation to the construction of a thesauri.

C3.1 When developing a thesaurus the identification and selection of thesaurus guidelines should be made at a very early stage in the process. Below are a few of these guidelines:

U.S. Office of Education, Rules for Thesaurus Preparation

UNESCO Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval

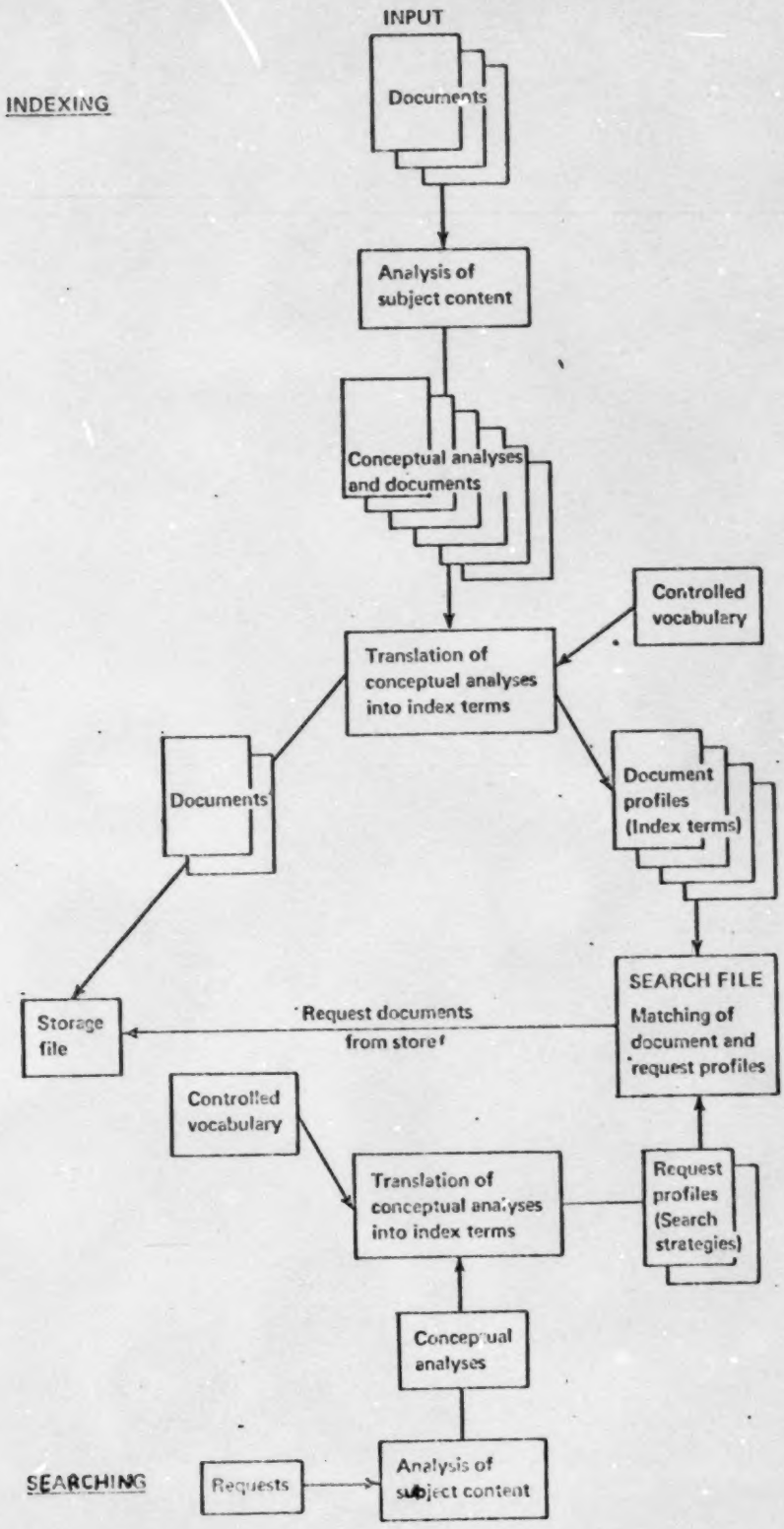


FIGURE 2 Information retrieval: input and output processes.

CHARACTERISTICS OF TYPES OF VOCABULARY MANAGEMENT TECHNIQUES

Wall , Eugene
Vocabulary Management - A Technical Report

Vocabulary Control Technique	Performance Range Encountered	CAPABILITY OF SOLVING BASIC PROBLEMS					COSTS		ELAPSED TIME REQUIRED	
			Semantic							
		Syntactic	Synonyms	Homographs	Generic	Viewpoint	Developmental	Operational (Maint. & Use)	Developmental	Operational Responsiveness
1. No vocabulary control; analyst selects terms from text.	Worst	Poor	None	None	None	None	Cheap	Expensive	Very Short	Quick
	Best	Fair	None	None	None	None	"Free"	Moderate	Very Short	Quick
2. No vocabulary control; title/text computerized and searchable.	Worst	Fair	Very Poor	None	None	None	Cheap	Expensive	Very Short	Very Quick
	Best	Fair	Poor	None	None	None	"Free"	Cheap	Very Short	Very Quick
3. Authority list without cross-references; no index.	Worst	Fair	Very Poor	Very Poor	None	None	Cheap	Very Expensive	Moderate	Very Slow
	Best	Fair	Poor	Fair	None	None	Cheap	Very Expensive	Short	Slow
4. Authority list without cross-references; with index.	Worst	Excellent	Very Poor	Poor	None	None	Moderate	Expensive	Moderate	Slow
	Best	Excellent	Poor	Excellent	Very Poor	None	Cheap	Expensive	Short	Moderate
5. Authority list with synonymous cross-references; no index.	Worst	Fair	Poor	Very Poor	None	None	Moderate	Expensive	Moderate	Slow
	Best	Fair	Excellent	Fair	None	None	Moderate	Expensive	Short	Moderate
6. Authority list with synonymous cross-references; with index.	Worst	Excellent	Poor	Poor	None	None	Expensive	Expensive	Moderate	Moderate
	Best	Excellent	Excellent	Excellent	Poor	None	Moderate	Moderate	Moderate	Moderate
7. Single-place classification schedule with cross-referenced index.	Worst	Fair	Poor	Fair	Poor	None	Expensive	Expensive	Long	Moderate
	Best	Excellent	Excellent	Excellent	Fair	Very Poor	Expensive	Moderate	Long	Quick
8. Taxonomy -- multi-place classification schedule with cross-referenced index.	Worst	Fair	Fair	Fair	Poor	Fair	Expensive	Moderate	Moderate	Quick
	Best	Excellent	Excellent	Excellent	Good	Good	Expensive	Moderate	Moderate	Quick
9. Information retrieval thesaurus.	Worst	Fair	Fair	Good	Fair	Good	Very Expensive	Moderate	Very Long	Quick
	Best	Excellent	Excellent	Excellent	Excellent	Excellent	Expensive	Moderate	Long	Quick

Each technique's capability of solving the basic problems is indicated on a scale extending from "none" to "excellent," as follows: None, Very poor, Poor, Fair, Good, Excellent.

The cost of implementing, maintaining, and using each technique is indicated on the following scale: Very expensive, Expensive, Moderate, Cheap, "Free".

The elapsed time for vocabulary control mechanism development is indicated on the following scale: Very long, Long, Moderate, Short, Very short.

The operational responsiveness of the vocabulary control mechanism with respect to timeliness is indicated on the following scale: Very slow, Slow, Moderate, Quick, Very quick.

ANSI, American National Standard Guidelines for Thesaurus Structure, Construction and Use

"National indexing standards have been developed for a number of reasons: to expedite the accurate indexing and searching of materials; to set guidelines for creating usable vocabulary products; to minimize the format differences in languages so that constant user retraining is unnecessary; and to facilitate the computer compatibility of indexing systems." The following works are also helpful in the construction process:

.Aitchison, Thesaurus Construction
.Gilchrist, The Thesaurus in Retrieval
.Lancaster, Vocabulary Development Control
.Soergel, Indexing languages and Thesauri Construction and Maintenance

If guidelines are chosen and followed they will assist in: 1) maintaining consistence of terms 2) avoiding proliferation of terms 3) clarifying ambiguities between terms.

C3.2 Identify area: An analysis of the literature may be conducted to determine the subject parameters of the discipline. This should include primary and secondary areas. In addition it should include linguistic factors.

C3.3 Selection of Term Groups. They are usually very broad grouping that most of the terms will fall under. They frequently are given code numbers.

C3.4 Selection of Terms. Once the main subject groups are clarified the collection of terms for these areas may begin. (In Lancaster)

C3.4.1 Manual selection

(A) Background sources
Background sources should be scanned for terms. Soergel has proposed this comprehensive list:
(i) Descriptor lists and thesauri.
(ii) Classification schemes (these include universal classification schemes such as Library of Congress and the Universal Decimal Classification) or parts thereof, and special classification schemes.
(iii) Nomenclatures of single disciplines.
(iv) Treatises on the terminology of a field or subject field.
(v) Encyclopedias, lexicons, dictionaries (user or discipline oriented, mono-bi-or multi-lingual).
(vi) Indexes of journals and abstracts journals.
(vii) Indexes of other publications in the field.

C3.4.2 Machine selection. In machine selection, lists of words are derived automatically from the text by using techniques such as statistical association. Most frequently used terms and also terms most frequently co-occurring are selected. Holloway states a conceptual weakness inherent in automatic indexing: "Automatic derivative indexing ignores the viewpoint expressed in a document." Despite the fact that Holloway was discussing the indexing process, her remarks are also applicable to machine selection to select terms. Even with this shortcoming these list can help the process of term selection.

C3.4.3 Usually a combination of these approaches are used. As Walls states: "Regardless of how terms are generated, for clarity they must be defined and limited in scope either by using existing lists, the consensous of specialists...or by arbitrary decisions."

C3.5 The decision of whether to include identifiers in the thesauri. Identifiers are names of people, geographical locations, organizations, projects etc. They are not structured and they do not usually appear in the thesaurus.

C4 Organization of Terms

Actchison notes that "a thesaurus differs from an alphabetic authority list of controlled terms in its ability to show structual relationship's between terms, hierarchical or non hierachical, as well as equivalence and associative relationships." The advantages frequently cited in the literature of developing a thesaurus in an information retrieval system relate to its underlying analytical, classified nature. Among these advantages are that a thesaurus

- Aids an indexer
- Aids the unknowledgeable searcher or searcher unsure of his specific need.
- Aids in rationally adding new concepts to the vocabulary.
- Allows for the expression of polyhierarchical relationships.
- Through its generic structure (BT) reduces the number of searches necessary to answer a general question. (Holloway)

The secret to these advantages lies in the organization of terms.

C4.1 The purposes of controlling syntactical problems are outlined by the Draft British Standard:

- To refer forbidden terms to terms that are synonymous with them.
- To form quasi-synonyms (also called near synonyms) into equivalence groups i.e. roughness/smoothness, altitude/height.

(B) Literature scanning

At the same time as terms are being selected from background sources, the literature should be scanned for terms. The indexing should not be limited to titles, but should include terms from the abstracts and conclusions. Soergel suggests that the indexing would be most effective if done by a number of different subject experts in the same field using terms of their choice.

(C) Question scanning

The terms arising from questions likely to be put to the system are as important as those taken from the literature. It is uneconomic to load a system with index terms which will not be required in searching. The questions should be collected from users or from records of questions already encountered.

(D) Individual experience and knowledge of compiler

The compiler's knowledge of the subject and familiarity with the terminology is an asset in term selection and rejection. But is would be unwise for the compiler to rely entirely on his own knowledge and memory of index terms without supplementing this with the scanning of question, literature and background sources.

(E) Individual experience and knowledge of the users

The compiler should also obtain the cooperation of users and user groups throughout the project. Compilation through a committee of experts may not be very satisfactory, but advice can be obtained in a more informal manner. It it suggested that:

(i) Experts may be asked to assist in the scanning of the literature, indexing items using the terms they consider most appropriate.

(ii) Experts may be asked to list terms of importance in their subject fields.

(iii) Experts may be asked to list typical questions which they might put to the system.

(iv) Experts may be shown lists of terms or draft classification schemes, in their own subject fields, and asked to comment, making amendments and adding terms.

(v) Groups of experts may be asked to discuss the terminology and classification in subject areas, which appear difficult to delineate and require careful clarification.

. To indicate preferred spelling
. To expand acronyms and abbrievations
. To indicate preference for concepts that are antonyms.
. To relate current jargon to more traditional terms (or vise versa).
. To express a concept by combining two or more terms in the manner of semantic factoring.
. To refer multi-term entries to a preferred word order
. To include translations of descriptors
. Aichison adds the following:
. The control of nouns and adjectives
. The use of singular and plural forms

C4.2 These are represented in the thesaurus by use and used for. Semantic problems have two aspects. Wall states:

. Synonyms exist when two terms have the same referend; different people may index the same information, and different people may search for the same information, using two or more different but synonymons terms.

. Homography occurs when a single word refers to two or more different referents. In this situation a parentical qualifier or scope note is used.

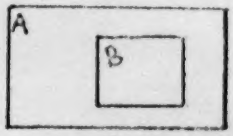
C4.3 The generic problem is associated with the broadness or narrowness of viewpoints of the originators of information, the indexer, and the searchers for information.

. Hierarchical Relationships. True hierarchical relationships include genus/specicies, and thing/type relationship. The part/whole relationship is sometimes taken to be a hierarchical relationship. The ESC Rules disallow this, except in the case of geographical locations, France/Paris. This is illustrated in the thesaurus by use of Broader Term (BT) and Narrower Term (NT)

Broader Term - Narrower Term Relationship (ERIC Processing Manual)

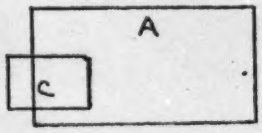
In information systems Descriptors (or terms) are names (or tags, or addresses) for classes of information, and a NARROWER TERM of a given Descriptor is simply a subclass within the scope of that Descriptor, as illustrated below:

Example 1:



If A and B represent terms in the Thesaurus, then Term B is a NARROWER TERM to Term A, because all members of the class represented by Term B are also members of the class represented by Term A. The corresponding reciprocal statement is, of course, that Term A is a BROADER TERM to Term B because, while it has members which are not members of Term B, it does include all members of Term B. In Example 2, on the other hand, Term C is not a NARROWER TERM

Example 2:



to Term A, nor is Term A a BROADER TERM to Term C, because although the terms have members in common, all of the members of Term C are not also members of Term A.

It is an invariable rule that a reference from a given term upward to a BROADER TERM must be reciprocated by a NARROWER TERM reference downward from the BROADER TERM, and vice-versa.

. Non-hierarchical related terms. The related term is one that is closely related conceputualy but not hierarchically. Willetts states:

> . It defines the descriptor by creating a set of conceptual relationships in the mind of the user
> . It alerts the user to other descriptors that might be helpful to him
> . It clarifies scope and defines specificily of the descriptor by putting it into a context
> . It supplements and compliments the classification schemes by connecting hierarchies
> . It links together terms that have close conceputual relationships
> . It is represented in the thesaurus by Related Terms (RT)

C4.4 The problem of viewpoint is relevant when different individuals view a given concept as belonging to different broad classes. This is resolved in a thesaurus by the use of parenthetical qualifiers, scope notes and generic relationships.

C5 Specificity and Multiword descriptors

"The specificity of a retrieval language depends on the ability of the index terms to express the subject in depth and in detail.

Specificity has an important influence on the performance of language as it constricts the accuracy with which concepts may be defined and consequently the facility to exclude unwanted items." (Aitchison)

The primary advantage of a highly specific vocabulary is that it allows the searcher to pinpoint the specific information that he wants and by doing so disregard unwanted citations. The disadvantage is that highly specific vocabularies are more expensive to compile, maintain and operate. The construction of a highly specific vocsbulary also calls for a more rigorous analysis of the subject area.

"Changes are needed more frequently in a specific vocabulary than in a system using only the more static concepts as index terms."

- To reduce effort and expense, specificity may be dropped to a lower level, and recall will not be lost if lead-in terms are made from the specific concepts to the broader terms used to represent them.
- In the same system different levels of specificity will be appropriate; a high level of specificity for areas central to the subject field and lower levels in the intermediate and marginal areas. (Aitchison)

Multiword terms (Bilingual Education) insure high precision because thev are more specific and exact. They however add to indexing costs by inflating the size of the vocabulary and they cause recall failures. Synthesized terms on the other hand (Bilingual and Education) give better recall although they can not give as high precision because they are less specific. The Rules for Thesaurus Preparation shown below defines when to use multiwords.

> Multiword descriptors (bound terms, precoordinated terms, and others) should be used whenever single-word descriptors cannot describe a concept adequately or provide effective retrieval. The following points should also be considered:
>
> > Use of a multiword descriptor is justified if any of the individual words in the multiword descriptor can combine so frequently with other descriptors as to produce many false coordinations.
> >
> > Use of a multiword descriptor to represent a unique concept is justified if the individual words of that multiword term are also unique descriptors which, when coordinated with each other, represent concepts different from the one intended by the multiword term.

Example: STUDENTS
TEACHERS
STUDENT TEACHERS

If a single-word term (used as a substantive) is so general as to be virtually useless in searching (e.g., SCHOOLS), consider the use of that term with another term (e.g., SECONDARY SCHOOLS).

Multiword descriptors, like single-word descriptors, must be carefully considered for placement in descriptor hierarchies.

Do not use inverted entries.

Examples: DEVELOPMENT
EMOTIONAL EDUCATION
ADULT
are not valid descriptors;

EMOTIONAL DEVELOPMENT ADULT EDUCATION
are valid descriptors.

C6 Recording of Terms. In order to keep tract of terms selected and their relationship it is customary to prepare a descriptor form for each descriptor to be included in the vocabulary. Appendices A & B

D Presentation of the Thesaurus. Most thesauri have the following layouts.

- Alphabetical
  - Index term
  - Synonyms, near synonyms & alternative word forms
  - Broader Terms
  - Narrower Terms
  - Non hierarchically related terms
  - Source (if taken from other vocabulary list)
  - Scope notes
  - Broad subject group or class number
- Rotated (Permuted) Display
- Hierarchical Display
- Descriptor Group Display

D1 Division of the Thesaurus (ERIC Processing Manual)

To assist the user (searcher or indexer) of the Thesaurus, the terminology is arranged in four different sequences, each providing a different approach. The principal arrangement is a complete alphabetic display of all Descriptors, with the other displays acting, in effect, as indexes.

PRIMARY ARRANGEMENT	SECONDARY ARRANGEMENT
Alphabetic Display	Rotated Display Hierarchical Display Descriptor Group Display

D2 Alphabetic Display

All Descriptors and USE references appear in this display interfiled alphabetically at one location letter-by-letter. Each Descriptor is shown with its complete record: hierarchical structuring, other cross-references, group code, and scope notes. All cross-references are reciprocal: for each USE there is a corresponding USED FOR: for each Broader Term there is a Narrower Term; for each Related Term there is a reciprocal Related Term.

Users who wish to determine the status of a certain term should first consult the Alphabetic Display. If a term is found, its meaning or definition can usually be determined by examination of the Broader Terms, Narrower Terms, Related Terms, or other cross-references appearing with it. Whenever necessary, however, a Scope Note is used to supplement the other details of the entry. A Scope Note is provided whenever there might be some doubt about the meaning of a term, and in order to ensure consistency of Descriptor usage. Scope Notes specify the range of subject matter to which a Descriptor is applied, draw necessary distinctions between closely related Descriptors, define obscure or technical terms, and state which one of several possible meanings is correct.

D3 Rotated Display

The Rotated Display sorts all single and multi-word Descriptors, considering each separate word as a filing unit. A Descriptor appears, therefore, in as many locations in this display as it contains separate words. The Rotated Display provides a permuted alphabetic index to all significant words that make up Descriptors in the Thesaurus (omitting USE references). Word order within Descriptor is not affected, e.g.:

ABILITY
ACADEMIC ABILITY
ABILITY GROUPING
LOW ABILITY GROUPING

This index tends to re-group related terms that may be separated in the alphabetic display due to word position within the descriptor.

D4 Hierarchical Display

The Hierarchical Display presents an alphabetic array of all the "Broadest Terms", i.e., all the terms that carry NT cross-references under them, but no BT cross-references above them. Terms having no BT-NT structure at all are also considered "Broadest Terms". Under each "Broadest Term" the particular generic tree involved is carried to its farthest extreme. Generic subordination (or class membership) is indicated via indentions:

LANGUAGES

.AMERICAN INDIAN LANGUAGES

. . . ATHAPASCAN LANGUAGES

. . . . NAVAHO

Terms which are narrower to some other term appear in this display only within that hierarchy, and not by themselves as separate alphabetic entries.

D5 Descriptor Group Display

Each of the Descriptors in the ERIC Thesaurus has been assigned to one (and only one) of 52 broad categories. The Descriptor Group Display shows each of these categories together with alphabetic arrays of the terms that have been assigned to each. The display is provided for browsing and easy scanning of large numbers of Descriptors that may be of interest within a given area. It is meant to provide suggestions and to be of general assistance to those using the Thesaurus for indexing, searching, and the structuring of new Descriptors. It is not a formal classification scheme.

Each Descriptor Group is defined by a Scope Note and is represented by a three-digit code number.

E Production of the Thesaurus

E1 "The computer can play several important roles in the handling of thesaurus data.

- It can manipulate and edit the input data, generate reciprocals, sort and print the thesaurus.
- It facilitates thesaurus updating and maintenance
- In a computer-based system, the thesaurus will be available in machine-readable form to allow certain mappings and term substitution to be conducted automatically, and provide statistics useful in indexing, searching, and vocabulary control functions." (Lancaster)

E1.1 The following computer packages have been identified that would assist in the production of the thesauri:

- Owens-Illinois, Inc
  Available from Aspen Systems
- AVOCON
  Available from Infodata Systems
- ERIC Thesaurus
  ORI
  Information Systems Division

All of the above are systems that are operated in a batch mode. There is one program developed by Vort and documented by Kazlauskas that allows you to develop a thesaurus on line (versus batch).

F Evaluation

F1 Checking with Subject Experts

The thesaurus should be reviewed by various subject experts. In order to do this more efficient the experts should be asked to comment on selected topics.

F2 Testing. Use the thesaurus to index a selection of documents at least 1,000 documents should be indexed before the terminology is finalized. Sample searches should be formulated and run against the system.

G Updating

All thesauri have to be updated since subjects are continously evolving.

G1 In order for the thesaurus to remain consistent a panel of information specialist, subject experts and a lexicographer a review process similiar to ERIC's should be formulated.

G2 Changes in the vocabulary fall under the following areas:

- lead in to broader terms
- synthesizing new concepts
- enumeration of new concepts
- deletion of terms
- revision of uncorrect or changing relationships

G3 Machine updating should be available through the computer aids discussed in section E.

H Guidelines for use of the thesauri must be developed. The chapter in The ERIC Processing Manual titled: Indexing should be used.

I Training (either formal or informal) must be conducted to insure consistent indexing. ERIC has recently finish a workbook entitled ERIC Abstractor & Indexer Workbook that serves as a model for training.

J Since the sixties many thesauri have been developed. These however have been mostly in the sciences. Nina Walls discusses some of the reasons for this: "Some obstacles to retrieval of information in the social sciences indicated in the forward to Foskett monograph on documentation in this area are the "imprecision of social science vocabularies, the proliferation of social science terminologies, the marked cultural and ideological contrasts in basic approaches to research." An additional problem is the overlapping between specialized subject areas. The situation in the humanities is probably similar, but the literature concerning documentation is even scantier." Commenting on the overlapping between areas Holloway states "The basic problem is that information seekers with different viewpoint may define the same concepts with different terms."

Holloway discusses Kuhn's theory: "Lodahl & Gordon discuss Kuhn's paradigm theory, the essense of the paradigm being... the degree of consensus or sharing of beliefs within a scientific field about theory, methodology, techniques, and problems. ...However Kuhn suggests that relative to the state of development there is a continum of fields with low development, such as political science or sociology, to those with a highly developed paradigm, such as physics. The state of development of a paradigm effects not only research but also, of interest to us, vocabulary control. The difficulty of defining a field and its vocabulary varies greatly from field to field dependant on the developmental state of its paradigm."

Hispanic Studies, not only comes from overlapping areas (many of which have no vocabulary control) but the paradigms which are used to describe the field are still in the development stages. A paradox to this situation is that the vocabulary can develop as does the field and may even aid in the development of the paradigm.

J1 In addition to the subject factors, linguistic concerns need to also be addressed. It may safely be assumed that a high percentage of the Hispanic origin individuals who will be using the data base are fluent in spanish. About 80% of the material is written in English. This presents us with the dilemma that if a user is monolingual spanish the subject approach in spanish will not significantly assist him if the material is in English. The answer, may be to make primary access to the data base in English, with terms commonly expressed in Spanish i.e. curanderismo, espiritismo be expressed in Spanish. Additionally, a Spanish version of the thesaurus could be developed after the system is in the operational stage.

good point

J2 As stated above controlled vocabularies outside the sciences are scarce, however a search of existing thesauri that are relevant to the Hispanic Bibliography project identified the following vocabularies: (See attachment C)

1. Library of Congress Subject Heading List
2. Thesaurus of Engineering and Scientific Terms (TEST)
3. Thesaurus of Psychological
4. Thesaurus of ERIC Descriptors
5. Hispanic Bibliography II Vocabulary
6. Chicano Thesaurus, for indexing Chicano material

K New Thesaurus versus Adapation

"Blagden asserts that the decision to construct a thesaurus for a subject field presupposes the inadequacy of existing lists of terms or their lack of suitability for a specific user group." The options are:

K1.1 Adoption of existing vocabulary with no modification. Most university libraries have adopted the Library of Congress subject heading list.

K1.2 Adoption of existing vocabulary as a framework, within which selected areas must be developed in greater detail. The Womens Educational Equity Communications Network has adopted this approach in utilizing the Thesaurus of ERIC Descriptors.

K1.3 Adoption of several specialized thesauri. The NTIS data base utilizes 10 thesauri to index their materials.

K1.4 The development from scratch of a new thesaurus.

K2 The relevant thesauri identified above will be briefly reviewed with regard to the options mentioned above.

K2.1 Library of Congress Subject Heading List. "The most widely used for subject analysis in academic libraries is the Library of Congress Subject Headings. These subject headings are the most widely used of the controlled precoordinated vocabularies primarily because of their presence on bibliographic records disseminated through the varied Library of Congress cataloging services directly to individual libraries or indirectly to them through commercial processing services and networks. Quantitatively this widespread use is impressive, but qualitatively it should not be interpreted as reflecting general approval of the headings. Numerous inconsistencies, anomalies, and other inadequacies have been documented and widely publicized by a number of critics. Daily identified problems in the structure of of the subject headings; Lilley and Taube were concerned about specificity. Perreault and Sinkankas identified the syndetic

Attachment C

	Scope Notes	Parenthic Qualifiers	Identifiers	Use	BT	NT	RT	Alphabetical listing	Rotated display	Classifitory aids hierarchical display or classifitory list	Date	# of Terms
LC	Yes	No	Yes	Yes	No	No	Yes	Yes	No	limited	1975	*
PA	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	1977	3500
TEST	Yes	Yes	limited	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1967	18000
ERIC	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1977	5200
Hispanic	Yes	No	Yes	No	No	No	No	Yes	No	No	1978	300
Chicano	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	1979	2000

* information not available

	Syntactical	Sematic synonyms	Sematic homographs	Generic	Viewpoint (scope note & related terms)
LC	Yes	Yes		No	Yes
PA	Yes	Yes		Yes	Yes
TEST	Yes	Yes		Yes	Yes
ERIC	Yes	Yes		Yes	Yes
Hispanic	limited	No		No	Scope note only
Chicano	Yes	Yes		Yes	Yes

structure as a serious problem. Berman has identified terminology as another basic problem, Berman, Clack (a), and Cazort have called for reforms in the subject headings for black Americans, while Marshell has pointed out the need for changes in subject headirgs relating to women." Haro and Cabello have pointed out "... a very low priority has been given to adequate indexing and the development of subject headings appropriate to Spanish Speaking groups."

Librarians, Researchers, students and others who search for Hispanic material on a regular basis are aware of the lack of specificity and relevance in the terminology used to describe Hispanic materials. This inadequacy presents two basic problems:

a. The inability to retrieve specific documents on specific aspects of the Hispanic Experience (precision)

b. The inability to retrieve all items that deal with the Hispanic Experience. (Recall)

Clack (a) states "... Not only are subject headings lacking in specificity and relevance when applied to black resources, but their unpredictable morphological and syntactical structure makes their usefulness less than ideal for subject searching", (these problems are addressed by thesauri). In summary, LC Subject Headings are a pre coordinated list which does not show hierarchical relationships. It has been documented as lacking in specifically and relevance for Minorities and Hispanics. It can be hypothesized that its generalized nature along with its inability to keep up with current terminology is part of the reason for the proliferation of specialized vocabularies. For the reasons outlined above the LC Subject Headings should not be used.

K2.2 Thesaurus of Pyschological Terms. This controlled vocabulary has all of the characteristics of a thesaurus. This thesaurus is used to index all the materials in Psychological abstracts. It however only covers psychology in depth and social sciences and the humanities marginally.

K2.3 Thesaurus of Engineering and Scientific Terms. This controlled vocabulary also has all the characteristics of a thesaurus. It is one of the thesauri used in NTIS. It's main emphasis is science and technology and only generally covers other areas.

Thesaurus of ERIC Descriptors.
This controlled vocabulary also has the characteristics of a thesaurus. This thesaurus was developed by the ERIC system which is a decentralized system with 16 clearinghouses. It has several terms which pertain to Hispanic Studies. In fact two of its clearinghouses have indexed between 6 and 7 thousand citations on Hispanics. This Thesaurus is one of the only thesauri that comes close to handling the areas of social science and humanities. As one would expect the greatist specialization

is in education. The thesaurus was based on both the literature and experts in the field of education. Additionally the hierarchical relationships are built around education concepts and do not reflect areas outside of that area adequately. The documentation and procedure for the development of this important thesauri are excellent and should be consulted.

K2.4 Hispanic Bibliography 11 vocabulary.
This is not a thesauri because it does not show generic, syntactical, or affinitive relationships. It does cover Hispanic populations such as Puerto Ricans, Mexican Americans and Hispanic concepts such as espiritismo and cuanderismo. It however has only 300 terms which is hardly enough to describe the whole area of Hispanic Studies. The emphasis is also psychological.

K2.5 Chicano Thesaurus, for indexing Chicano material.
This controlled vocabulary was developed by first combining five different Chicano Subject Heading lists in a single subject heading list. After a review by Kocalis it was decided to restructure the list into a thesaurus using the ERIC guidelines. This restructuring has only recently been completed. Although the primary Hispanic groups are represented in the thesaurus not all Hispanic concepts are represented i.e. espiritismo. The thesaurus is also more developed in some areas i.e. Folklore, labor than in others Psychology.

K3 Although none of the thesauri reviewed completely provides both structure and terminology for the area of Hispanic Studies the Chicano comes the closest both in terms of providing a baseline terminology and in terms of structure.

K3.1 Based on the options presented in J2 it is provitionally recommended that the Chicano Thesaurus be used as the baseline document within which selected areas be developed in greater detail. Final recommendations will be made after review of the Thesaurus by a selected group of Hispanic Scholars.

K3.2 Revision of the Thesaurus, if adopted, will require the following activities:

- Selection and definition of the areas that will be covered in the thesaurus.
- An indepth review of the Thesaurus by a group of scholars and librarians in the area of Hispanic Studies.
- The adoption of Thesaurus Guidelines. (ANSI Guidelines are recommended).
- Review of the descriptors that appear in data bases that heavily posted with Hispanic material for possible encorporation into thesaurus. (a survey of these data bases is currently being conducted by our project)

- Indexing of a 1000 documents. This will serve as both an evaluation of the thesaurus and a way of generating new candidate terms.
- The formation and implementation of a procedure for reviewing candidate terms.
- Production of the thesaurus by one of the software programs identified in E.

L. Related Thesauri

APA
Thesaurus of Psychological Terms, Washington, D.C., American Psychological Association, 1977.

Ad Hoc Committee for Developing a Chicano Thesaurus
The Chicano Thesaurus, for indexing Chicano materials, Berkeley, Chicano Periodical Index Project, 1979.

Cox, Barbara G. & Larson, Leslie
Hispanic American Periodicals Index Subject List 1975, University of California, Los Angeles Latin American Center, 1977.

ERIC
ERIC Thesaurus of ERIC Descriptors, New York, Mac Millian, 1977.

U.S. Library of Congress
Library of Congress Subject Headings, Washington D.C., Library of Congress, 1975.

Padilla, Amado et al
Hispanic Bibliography II, Los Angeles, Spanish Speaking Mental Health Research Center, 1978. Includes list of descriptors.

Documentation Associates
National Clearinghouse on Aging-THESAURUS, (July 1977), 2nd Edition, Washington D.C., U.S. Department of Health, Education and Welfare, Office of Human Development Services. Administration on Aging National Clearinghouse on Aging, 1977.

Rosenberg, Paul M.
The Urban Information Thesaurus , A Vocabulary for Social Documentation , Greenwood Press, 1977.

U.S. Department of Defense
Thesaurus of Engineering and Scientific Terms, Washington, NTIS, (1967) 1969.

M. Selected Bibliography

ANSI
"American National Standard Guidelines for Thesaurus Structure, Construction, and Use," New York, American National Standards Institute, Inc., 1974.

Aitchison, Jean & Gilchrist, Alan
Thesaurus Construction, A Practical Manual, London, Aslib 1972.

Berman, Sanford
"Getting to It; a Guest Editorial", American Libraries, Vol 8 (Feb 1977) pp. 77.

Berman, Sanford
Prejidices and Antipathies: A Tract on the L.C. Subject Headings concerning people. Metuchen, N.S., Scarecrow Press, 1971.

Block, Margaret
Analysis of Existing Controlled Vocabularies: Women's Educational Equity Communications Network, U.S. Department of Health, Education and Welfare, 1978.

Booth, Barbara
VIP Manual, Guidelines and Procedures for ERIC Vocabulary Improvement, ERIC, December 1978.

Blagden, John
"Structured Thesauri", Aslib Proceeding, Vol 23, No 3, (March 1971) pp. 139-143.

Blagden, J. F.
"Thesaurus Compilation Methods: A Literature Review", Aslib Proceeding, Vol 20, No 8, (August 1968) pp. 345-359.

British Standards Association
Notes for a draft standard. Guidelines for the establishment and development of monolingual thesauri for information retrieval, London, 1972.

Cabello - Argandona, Roberto and Roberto Haro
Systems Analysis of Library and Information Services to the Spanish Speaking Community of the United States, Los Angeles, Chicano Studies Center, 1977.

Cazort, Jean Elder
A Handbook for the Organization of Black Materials. Prepared for the Institute on the Selection, Organizations and Use of Materials by and about the Negro, Nashville, Tenn., Fisk University, 1971.

Chicano Affairs Program
Editorial Chicano, Stanford, Chicano Affairs Program, 1978.

Chicano Studies Center
Guide to Chicano Studies Departments, Programs, and Centers, Los Angeles, Chicano Studies Center, 1975.

Chicano Studies Center
National Directory of Chicano Faculty and Research, Los Angeles, Aztlan Publications, 1974.

Clack, Doris Hargrett
"The Adequacy of Library of Congress Subject Headings for Black Literature Resources", Library Resources & Technical Services, Vol 22, No 2, (Spring 1978) pp. 137-144.

Clack, Doris H.
"Analysis and Organization", Black Literature Resources, New York, Decker, 1975.

Clack, Doris H.
"An Investigation into the Adequacy of Library of Congress Subject Headings for Resources for Black Studies", unpublished dissertation, School of Library and Information Sciences, Univ. of Pittsburgh, 1973.

Cosati
Guidelines for the Development of Information Retrieval Thesauri, U.S. Federal Council for Science and Technology, Committee on Scientific and Technical Information (COSATI). Washington, D.C., Government Printing Office, 1967.

Daily, Jay E.
"The Grammer of Subject Headings; a Formulation of Rules for Subject Headings Based on a Syntactical and Morphological Analysis of the Library of Congress List", unpublished dissertation in School of Library Service, Columbia Univ., 1957.

ERIC
"Vocabulary Development And Maintenance", Eric Processing Manual, ERIC, July 1, 1974, pp. 257-353.

EJC
Thesaurus Rules and Conventions. New York, Engineers Joint Council, Nov 1966.

Foskett, A. C.
The subject approach to information, Wales, Archon Books and Clive Bengley, 1969.

Gilchrist, Alan
The Thesaurus in Retrieval, London, Aslib, 1971.

Goodman, Frederick Dr.
"The Role and Function of the Thesaurus In Education", Journal of Library Automation, Vol 2, No 4, (December 1969).

Holloway, Dona
"Work On The Theoretical Foundations of Vocabulary Control", Student Paper, School of Library and Information studies, University of California, Berkeley, (Spring Quarter, 1978), pp. 1-18.

Kazlauskas, Edward John
Evaluation Of The Spanish Speaking Mental Health Research Center Bibliographic Data Base Vocabulary, January 8, 1979, pp. 1-12.

Kocalis, Elva Aguirre
"Diversity of materials to be indexed by the Chicano Periodical Index Project", Report for the Chicano Periodical Index Project, University of California, Berkeley, 1978.

Kocalis, Elva Aguirre
"Recommendations for the Development of the Chicano Thesaurus for Use in Periodical Indexing", Report for Chicano Periodical Index Project, University of California, Berkeley, 1978.

Lancaster, F. W.
Vocabulary control for information retrieval, Wash., D.C., Information Resources Press, 1972.

Lilley, Oliver Linton
"How Specific Is Specific?", Journal of Cataloging and Classification, Vol 11, (Jan. 1955), pp. 3-8.

Lilley, Oliver Linton
"Terminology, Form, Specificity and the Syndetic Structure of Subject Headings for English Literature", unpublished dissertation, School of Library Service, Columbia Univ., 1958.

MacCafferty, Maxine
Thesauri & Thesauri Construction, London, Aslib, 1977.

Marshall, Joan K.
On Equal Terms: A Thesaurus for Nonsexist Indexing and Cataloging, New York, Neal-Schuman Publishers, 1977.

Meincke, Peter & P. Atherton
"Knowledge space: a conceptual basis for the organization of knowledge, JASIS, Vol 27, No 1, (Jan-Feb 1976), pp. 18-24.

Neufeld, Margaret L.
"Linguistic Approaches to the Construction and Use of Thesauri: A Review", Drexel Library Quarterly, Vol 8, No 2, (April 1972), pp. 135-208.

Office of Education Panel on Education Terminology (PET)
"Rules for Thesaurus Preparation", Washington D.C., U.S. Government Printing Office, 1969.

Perreault, Jean M.
"Re-Classification: Some Warnings and a Proposal", Urbana, Univ. of Illinois Graduate School of Library Science, 1967, pp. 7-8.

Pickford, A. G. A.
"Some Problems of Using An Unstructured Information Retrieval Language In A Coordinate Indexing System", Aslib Proceedings, Vol 23, No 3, 1971, pp. 133-138.

Rogers, Virginia G.
"Thesaurus Construction: An Introduction", Drexel Library Quarterly, Vol 8, No 2, (April 1972), pp. 117-124.

Sinkankas, George M.
A Study in the Syndetic Structure of the Library of Congress List of Subject Headings, Pittsburgh, Graduate School of Library and Information Sciences, 1972.

Soergel, Dagobert
Indexing languages and thesauri; construction and maintenance, Los Angeles, Melville, 1974.

Taube, Mortimer
"Specificity in Subject Headings and Coordinate Indexing", Library Trends, Vol 1 (Oct 1952), pp. 219-223.

UNESCO
Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval. Paris, United Nations Educational, Scientific and Cultural Organization, July 1970.

Vesenyi, Paul E.
"An Experiment with Bilingual Subject Headings", The Indexer, Vol 10, No 2, (October 1976), pp. 73-74.

Wall, Eugene
"Symbiotic Development of Thesauri and Information Systems: A Case History", Journal of The American Society for Information Science, (March-April), pp 71-79.

Wall, Eugene
Vocabulary Management - A Technical Report, Germantown, Maryland, Aspen Systems Corporation, N.J., N.d.

Walls, Nina
"Thesauri and Classifications in Non-Scientific Areas", Drexel Library Quarterly, Vol 8, No 2, (April 1972), pp. 129-134.

Willetts, Margaret
"An Investigation of The Nature of The Relation Between Terms In Thesauri", Journal of Documentation, Vol 31, No 3, (September 1975), pp. 158-184.

Appendix A

1A Descriptor Ident. Code	1B Descriptor – In Upper Case	1C Descriptor Origin
2A Descriptor in Final Print Format (SN)		
2B Subject Categories (SC)	2C Scope Notes (SN)	
3 Use or "Post On" Descriptor (US)		
4 Use for - Reciprocal Use (UF) Cross Reference		
5 Broader Descriptors (BT)	7 Related Descriptors (RT)	
6 Narrower Descriptors (NT)	8 Nonreciprocally Related (RR) Descriptors	
9 Date of Final Review	10 Date Added to Thesaurus Corpus	

FIGURE 39 Master descriptor form used for thesaurus input.

Appendix B

Thesaurus Term Review Form

THESAURUS TERM REVIEW

DATE SUBMITTED: DAY / MONTH / YEAR

SUBJECT TERM

ACTION REQUIRED

ADD AS A NEW TERM: POSTABLE ☐ NONPOSTABLE ☐

CHANGE TERM TO: POSTABLE ☐ NONPOSTABLE ☐ SPELLING FROM

CROSS REFERENCE: ADD ☐ DELETE ☐ REFERENCES SHOWN BELOW

DELETE TERM: UNLESS ONLY ☐ DELETE & TRANSFER ☐ POSTINGS TO

SOURCE DATA

ACCESSION NO

SUGGESTED BY

TERM USED IN LIEU OF REQUESTED TERM

AMPLIFYING INFORMATION

SCOPE NOTE

DEFINITION

CROSS-REFERENCE STRUCTURE

USE

USED FOR (UF)

BROADER TERMS (BT)

NARROWER TERMS (NT)

RELATED TERMS (RT)

COMMENTS

SUBMITTED BY

TITLE

REVIEW AUTHORITY

REVIEW COMMENTS

ACTION: APPROVED ☐ DISAPPROVED ☐

REVIEWER'S SIGNATURE

DATE: DAY / MONTH / YEAR